On October 13, Columbia University played host to a distinguished guest when Bill Gates, the Chairman and Chief Software Architect of Microsoft, visited campus. Gates visited Columbia as part of a six-university tour undertaken with the goal of fostering interest in computer science and related fields; this follows a similar tour of five other campuses last spring. Computer Science department chair Henning Schulzrinne stated that “It validates Columbia as one of the very small number of places in the U.S. that Gates put on his list of five or six places that he’s visiting... I think that’s a very good sign that Columbia is one of those places that somebody of his stature would want to visit.”

The Faculty Roundtable
Before speaking to a packed audience of students in Lerner Hall, Gates participated in a roundtable discussion with Computer Science department faculty. Department chair Henning Schulzrinne started things off with a brief presentation on “Columbia’s Vision for Tomorrow’s Global Intelligent Systems.” Henning gave a bird’s-eye view of the many different active research areas in computer science at Columbia, ranging across topics such as theory, digital systems design, computer systems and networks, interacting with humans, graphics/ robotics/ vision, and making sense of data. He highlighted several innovative research projects underway at CUCS, including the Newsblaster system for automatic summarization that has been developed by Professor McKeown and her team and the work on distributed channel allocation that is being carried out by Professors Misra and Rubenstein.

Following Henning’s presentation, Professor Shree Nayar gave an overview of his lab’s research on computational cameras and displays. Shree started off by talking about his research on wide angle imaging (which has led to commercial development of 360-degree cameras for videoconferencing and other applications) and on radial stereoscopic imaging, which enables a “depth camera” to generate three-dimensional images. Shree also described some exciting recent work using randomized point clouds which may soon lead to inexpensive three-dimensional displays constructed of laser-damaged glass. At the end of Shree’s presentation, he gave Bill Gates the actual glass cube that was used to implement a three-dimensional version of Pac-Man. (As Shree noted, this was one of the few things that Bill would have a hard time acquiring for himself.) Perhaps the high point of the faculty roundtable was a question-and-answer session between Gates and Computer Science faculty members. To start things off, Professor Kathy McKeown noted that (continued on next page)
2 How do we prove that certain problems are hard to solve?
3 How can we make truly reliable software? Is there something analogous to Shannon’s work on error correction and von Neumann’s work on reliability through redundancy?
4 How should we architect the systems so that they can be more easily maintained and evolved?
5 How can we make computers understandable and usable for people of all backgrounds, ages, and abilities, in the many different situations we encounter in life?
6 How can we program computers to have human qualities such as consciousness, intelligence and emotion?

At Gates asked Gates how Microsoft and universities might systematically work together to address these issues. In response, Gates stated that there is no easy solution and that the dropoff in women and minorities occurs at every stage of the pipeline—from early interest in math and science, through the study of computer science courses, selecting a major, and even to the profession itself. Just as there is dropoff at all these stages, he advocated working to increase the participation of women and minorities in computer science. To do this, Gates advocated starting early outreach at summer camps and high schools all the way through college and industry efforts. (As Gates briefly noted, Microsoft Senior Vice President Rick Rashid is a trustee of the Anita Boren Institute for Women and Minority Students.)

A second question from Professor Al Aho concerned the future of computer science. Al had previously canvassed the Computer Science faculty for their "top five" most important unsolved questions in computer science, and they came up with the following list:

1. How do we prove that certain problems are hard to solve?
2. How can we make truly reliable software? Is there something analogous to Shannon’s work on error correction and von Neumann’s work on reliability through redundancy?
3. How should we architect the systems so that they can be more easily maintained and evolved?
4. How can we make computers understandable and usable for people of all backgrounds, ages, and abilities, in the many different situations we encounter in life?
5. How can we program computers to have human qualities such as consciousness, intelligence and emotion?

At Gates asked Gates how Microsoft and universities might systematically work together to address these issues. In response, Gates stated that there is no easy solution and that the dropoff in women and minorities occurs at every stage of the pipeline—from early interest in math and science, through the study of computer science courses, selecting a major, and even to the profession itself. Just as there is dropoff at all these stages, he advocated working to increase the participation of women and minorities in computer science. To do this, Gates advocated starting early outreach at summer camps and high schools all the way through college and industry efforts. (As Gates briefly noted, Microsoft Senior Vice President Rick Rashid is a trustee of the Anita Boren Institute for Women and Minority Students.)

A second question from Professor Al Aho concerned the future of computer science. Al had previously canvassed the Computer Science faculty for their "top five" most important unsolved questions in computer science, and they came up with the following list:

1. How do we prove that certain problems are hard to solve?
2. How can we make truly reliable software? Is there something analogous to Shannon’s work on error correction and von Neumann’s work on reliability through redundancy?
3. How should we architect the systems so that they can be more easily maintained and evolved?
4. How can we make computers understandable and usable for people of all backgrounds, ages, and abilities, in the many different situations we encounter in life?
5. How can we program computers to have human qualities such as consciousness, intelligence and emotion?
**Feature Article**

**Women In Computer Science (WICS) at Columbia**

In our field a woman’s face is as hard to find as a fix for a buggy program that is due in an hour. WICS: Women In Computer Science (WICS) is a group that is changing this situation by bringing together women involved in computer science in the Columbia University community. WICS hopes to encourage prospective computer science majors and provide direction and resistance to current computer science students. "WICS also strives to increase the number of women undergraduates, grad students, faculty and staff, the group promotes interaction on social, academic and professional levels."

WICS hosts social events, corporate information sessions, and educational functions primarily catered to women but also open and often very relevant to men as well. Among WICS events are corporate information sessions, faculty panels that address topics relevant to undergraduates and graduate students such as “Giving Talks,” “Choosing Research Topics,” and “Job Options After Graduate School.” Naturally, one of the most important events that WICS is the forum it provides to address women’s concerns, questions, ideas and goals. To this end WICS organizes book discussions on topics specific to women such as “Don’t Ask: Negotiation and the Gender Divide,” by Linda Babcock and Sheela Karthik, and “Women in Science: Meeting Career Challenges,” by Angela M. Pattatucci, et al., and recently participated in a university-wide panel on Sexual Harrassment. WICS also facilitates access to mentorships and opportunities for students in SEAS. In a joint effort with Carnegie Mellon University, WICS organized a dinner with distinguished CMU professors and graduate students and faculty from NYU, Stevens, Hunter, and NJIT, funded by the CMU. The dinner was followed by a WICS-Distinguished Lecture given by Dr. Lenore Blum of CMU and lunch sponsored by CFA-W with a panel on graduate students moderated by Dr. Joann Orfale, Avaya Labs. The activities finished with a presentation by CMU students “Exploring Graduate School: The Women @SEAS Outreach Roadshow.”

WICS is currently planning a Fall welcome party, a student social, and a panel to address hidden barriers to success in academia, and a distinguished lecture by a local woman scientist. WICS’ activities are funded by the Computer Science Department, Google, and by the Vice-Provost for Diversity and Equity. WICS’ Executive Board includes faculty, research staff, graduate and undergraduate students. In summary, WICS is an active and visible voice in the computer science community that seeks to bring women with the opportunity to organize and participate in events seeking to improve the university experience as a whole.

For more information you can visit the WICS website at http://www.columbia.edu/wics or contact WICS by email at cu-wics@columbia.edu.
The Natural Language and Speech Processing Groups

Natural language and speech processing deal with the theory and practice of enabling computers to understand and produce language, both written and spoken.

Human-computer interfaces based on language understanding and generation have the potential to make online information universally accessible to the general public.

At Columbia, the Natural Language and Speech Processing Groups are addressing problems in analyzing and providing access to large amounts of information, whether spoken or written. We are developing browsing and summarization systems for news (both published written sources and broadcast news), email, journal articles, and meeting notes. Our research in spoken language processing also aims at identifying deceptive speech and emotions. We are also developing techniques for multilingual analysis, whether spoken or written.

The group includes Prof. Julia Hirschberg and Prof. Kathleen McKeeon, research scientists Martin Jansche and Oren Rabinow, post-docs Stefan Benus, Mona Diab, Nizar Habash, and Advaith Siddharthan, and consultant Rebecca Passonneau. 14 PhD students and two programme assistants are also part of the group.

Summarization

Our research addresses automatic summarization of a wide range of genres, from formal written language (e.g., medical journal articles) to informal, conversational speech (e.g., meetings).

Newsblaster

Current technology in summarization and topic detection and tracking is mature enough to be used reliably in a live online environment. A number of people in our group have developed Newsblaster, a system that crawls news sites, filters out news from nonnews (e.g., ads), groups news into stories on the same event, and generates a summary of each event. Newsblaster generates summaries by extracting important sentences and by generating new sentences that fuse information from different documents. News is automatically categorized into five main categories (US, World, Entertainment, Sports, and World), and within each category, news stories are automatically clustered into stories on the same event. It tracks events across days and can generate an update identifying what’s new on a given day in comparison to previous days. We have also developed a multilingual version of Newsblaster, which crawls foreign language news sites and generates English summaries of documents on the same event drawn from many languages. Newsblaster runs on a daily basis, crawling 25 news sites (see http://newsblaster.cs.columbia.edu). The system can be searched or browsed, to locate stories of interest, and these stories can subsequently be summarized for the user.

Broadcast News

In our work at Columbia summarizing Broadcast News, we have pursued a two-level approach to the problem of summarizing errorful spoken material. First, we identify domain-specific aspects of newscasts to provide an outline of the newscast, which users can navigate in a GUI interface, following links from e.g., headlines to stories and speakers to the speech they contribute. In domains like Broadcast News, the material to be summarized exhibits far more regular patterns from one speech document to another than news broadcasts generally open with a news anchor’s introduction of the major news stories to be presented in the broadcast. Followed by the actual presentation of those stories, and then the closing of the broadcast. The summarization system finds key questions posed within the thread and their corresponding answers, and integrates these with sentences that it determines contain important information. To find these sentences, it uses a machine learning approach based on features related to real networked users, with highest honors from Columbia and holds a Ph.D. in computer science from Carnegie Mellon, is a former corporate vice president at Microsoft who is now heading Google’s operations in China. A student under the supervision of Tom O’Donnell of Siemens Corporate Research (Ph.D. Program 1990) has won the Best Student Paper in the Area of Segmentation and Processing in this year’s MICCAI (Medical Image Computing and Computer Aided Intervention) conference. “Quantification of Delayed Enhancement Images”. By Engin Dikici, Thomas O’Donnell, Randolph Setser, and Richard D. White was awarded a cash prize of 500 Euros.

Alumni News

Bruce Abramson (BA ’83, MS ’85, Ph.D. ’87) writes: “My first book intended for a general audience, Digital Prophets: Why the Information Economy, Commerce and How to Win It Rise Again, (MIT Press, 2001), is now available through bookstores everywhere. Digital Prophets is my attempt to make sense of the major economic events of the past decade—in particular, the Internet investment bubble, the Microsoft trial, the rise of open source, and the P2P wars. In doing so, I draw upon some results that those of us well-versed in artificial intelligence, software engineering, industrial organization, network economics, antitrust, and intellectual property law may know well, but that remain news to most of the rest of the world. I also spend a chapter discussing the relationship between the information economy and the broader social and political are-

nas that are beginning to feel the transition from industrial age to information age. Despite that potential of intimidating pro-
file, I worked hard to make the material accessible to a general audience, and above all to make the book an entertaining read. I’ve collected early reviews on my website, at http://www.theinformationist.com/index.html#reviews/.”

Andrew Arnold (CC ’93) is in his second year of Carnegie Mellon’s Ph.D.program in Information Management. He reports that he is “enjoying Pittsburgh, but still pining for the city.”

Regina Barzilay (Ph.D. ’02) was selected as one of this year’s top 35 technology innovators under the age of 35 by Technology Review. She is currently an assistant professor at the Computer Science and Artificial Intelligence Laboratory at MIT. Regina is being recognized for her work in teaching computers to read and write; as the Technology Review article reports, “For her doctoral dissertation at Columbia University, computer scientist Regina Barzilay led the development of Newsblaster, which does what no computer program could do before: recognize stories from different news services as being about the same basic subject, and then paraphrase elements from all of the stories to create a summary.”

Jeannie Flesmer (Barnard ’96) is moving down to Washington, DC, where she will be clerk for Justice David H. Souter of the U.S.Senate.

Stuart Haber (Ph.D. ’86) is the General Chair of the 25th International Cryptology Conference (CRYPTO 2005), which is the main conference on cryptography held annually in the United States. Stuart was quoted in an August 17 article in the New York Times on crypto-
Conference every eighteen months. The group is chaired by Prof. Don Towsley, UMass.

Professor Steve Nowick has been brought onto the DARPA CLASS program, a major government initiative to translate asynchronous digital design viable for the commercial and military sectors. There were large-scale proposals submitted, and only one funded, headed by Boeing, with participation of Philips Semiconductors, two asynchronous startups and two smaller academic efforts. The two goals of the project are to build a large-scale asynchronous demonstration chip (for Boeing) and design an asynchronous CAD tool for use in design of asynchronous designs. Prof. Nowick and his former Ph.D student Moneek Singh (currently an assistant professor at UNC), will play a key role in transforming their high-speed asynchronous pipeline style, MOUSERAP to the flow, and providing optimizations for several of the major design flow steps.

PhD student Rafi Pelosof and senior research scientist Yow Freund, both at the Center for Computational Learning Systems (CLLS), presented their work on their shared co-located poster in a full day of the CVPR (Computer Vision and Pattern Recognition) conference, held in San Diego in June.


Professor Henning Schulzrinne was interviewed about the 2hub file sharing system by ABC Nightline and by Fox News. Professor Henning Schulzrinne won the 2005 Spunkinn Innovator Award for his contributions to VoIP. Prof. Schulzrinne received the award at the 2005 forward2business Conference in Halle, Germany.

Professor Henning Schulzrinne, along with a team consisting of researchers from Pennsylvania State University, University of California-Santa Barbara and Lucent Technologies, won a National Science Foundation grant titled WORKIT: A Universal Wireless Open Research Kit. The WORKIT project addresses the need for wireless network tools and platforms as recom- mended in the 2003 NSF: Wireless Network Workshop report. The project will build on the IOTA (Integration of Two Access Technologies) project and a distributable form called the Wireless Open Research Kit (WORKit). WORKit will include a source code and documentation and also be embedded in low-cost off the shelf hardware. WORKit will be an enabler for research in mobility management, interlayer awareness, software algorithms for optimal network selection, reconfiguration, authentication, policy definition, and performance, hybrid wireless networking. Broader impacts of this project include WORKit in education and enabling stronger universi- ty-industry collaborations in this area of emerging importance at colleges and universities.

Professor Rocce Servidio was elected as a board member of the Association for Computational Learning; the members of the board comprise the steering committee for the Conference on Computational Learning Theory (COLT).

Professor Rocce Servidio was awarded a grant from the NSF program on Emerging Models and Technologies for Computation (EMT). The EMT cluster seeks to advance the fundamental capabilities of computer and information science and engineering by capitalizing on advances and insights from areas such as biological systems, quantum phenomena, nanoscale science and engineering, and other novel computing concepts.

The award will support Rocce’s research on connections between quantum computation and computational learning theory.

Two DHS grants have been awarded to CounterStrom and Columbia’s IDS Lab, headed by Professor Saul Stofflo. This project aims to research and develop a new generation of collabora- tive, cross-domain security systems to prevent the exploitation of net- work-based computer systems. The core concept is to deploy a number of strategically placed sensors across a number of participating networks that collaborate by sharing information in real time. Each participating network and each of its mem- bers is a novel content-based anomaly detector, PAVI, identi- fies likely new exploits target- ing vulnerable systems. The Worminator project has developed a new generation of scalable, collaborative, cross-domain security systems that exchange alert information including pro- posed behaviors of attacks and privacy-preserving anomalous content alerts to detect severe zero-day security events. The work is a joint collaboration with CounterStorm, a New York City based company spun out from the DSI and DARPA- sponsored Columbia project.

Professor Joseph Traub was named as the new chair of the Computer Science and Telecommunications Board (CSTB) of the National Academy of Sciences. This CSTB deals with critical issues facing the nation in the area of computer science and telecommunications. Projects include cybersecurity research, biometrics, IT to enhance the independence and building certify depend- able systems. For more infor- mation visit the website of Prof. Traub’s appointment at Columbia University, as he was also its founding chair. “In 1986, along with Marjory Blumenthal, Joe’s vision and dedication established the model that has made CSTB one of the strongest boards at the Academies. At this particular point in CSTB’s history, one of the most pressing issues is how another person better suited to assume the chair and lead CSTB to new heights,” said Bill Wulf, President of the National Academy of Engineering.

Medical Journal Articles and Textbooks
We have developed a system to summarize journal articles, textbooks and online consumer health as part of the CounterStorm project being developed at Columbia University, called PERSIVAL. PERSIVAL (Personalized Retrieval and Summarization of Images, Video, and Audio) is a framework, which aims to provide tailored pres- entation of the relevant medical facil- ity for physicians and physi- cians and laymen. The PERSIVAL summarization component takes as input doc- uments relevant to a user’s query, retrieved by a search component. It generates an English summary of one or more paragraphs of the group of documents, highlighting facts that are common to all documents and pointing out differences between them. A key feature of the PERSIVAL summarization component is to personalize its content given information about patient sta- tus available in the online patient record; the result is a summary with knowledge that is more relevant to the user, where the physician treating the patient.

Meetings
Meetings involve multi-party conversations, and conversations with speakers; the language is informal and utterances tend to be short and superficial. One goal of the project is to develop MSA morphological and grammatical and many ellipses and pronouns. At Columbia, we are working on making summarization as part of the system project entitled Mapping Meetings where the goal is to create methods for efficiently browsing and visualizing meetings. Our work to date has focused on methods for demar- cating utterances, the most important content and for gener- ating summaries of a summary. We have developed techniques for segmenting the meetings into different topics for recognizing agreement and disagreement in dialog, and are currently working on methods for generating summary sentences that can remove redundancies and extraneous material.

Analyzing Spoken Language
Deceptive Speech
We are currently examining the feasibility of automatic detection of deception in spoken language, synchronous and asynchronous. Our current study consists of a constant in which the subject is positively motivated to deceive an inter- viewer, where deceptive and truthful statements are marked by the subject. The interviewee recognizes the content of the subject’s speech extracted for statistical analyses and machine learning experiments that attempt to find those fea- tures best able to distinguish true from lie.

Emotional Speech
This research seeks to identify acoustic and prosodic cues to emotion. Other CAD tools to subjective judgments of emo- tion as well as the objective eye-tracking studies of sub- jects asked to match audio data with human faces which display different emotional states. The goals of the study are to first establish those acoustic and prosodic features of an utterance which are most correlated with subjective judgments of emotion and then to determine the feasibility of automatic recognition of emotional states.

Machine Learning
Our current focus is on methods for identifying emotions in speech. In many cases given a set of audio samples, the task is to label the emotions expressed in the speech. In other cases, the given information is a chain of sentences. The goal is to be able to identify the emotion expressed in each sentence.

Charismatic Speech
People at an influential level are drawn to certain public speakers. What about it makes their speech charismatic? Our research examines acoustic and textual features of charis- matic’s speech in a variety of contexts to identify the acoustic, prosodic and lexical indicators that best correlate with subjects’ judgments of charisma. Though our work to date has been on American English, speech by Syrian and Palestinian Arabic is being conducted to identify cultural differences in the perception of charisma.

Arabic DIY Language Processing
Three members of the Natural Language Processing Group are part of the Center for Computational Learning Systems where they are concentrated on Arabic and its dialects. The Arabic language is actually a collection of hundreds of dialects, both phonological, morphological, lexical, and syntactic differ- ences comparable to those seen among the Romance lan- guages. Arabic is the most widely used language in the Arab world, the standard written language is the same. Modern Standard Arabic (MSA), which is also used in some official spoken communica- tion (newscasts, parliamentarian arguments). MSA is based on Classical Arabic and is itself composed of so many dialects. Other forms of Arabic (general- ly referred to as “dialects” of MSA) are what people use for daily spoken communication, in unofficial written communica- tion, in particular in the now growing electronic media, often ad hoc transcriptions of dialects are used. The bound- aries between MSA and a dialect are not well defined and borrowing between the two forms occurs often.

Within the Arabic-dialect family, MSA is the variant with most available resources in terms of corpora (mongolian text, English-Arabic parallel texts, and Treebanks), morphological ana- lyzers, etc. There are far fewer resources available for most other dialects; since they are mainly spoken, even unannotat- ed corpora are not common.

The focus of our current research on Arabic Dialects is to create NLP tools for Arabic dialects, including resources poor dialects. Our approach is to create frameworks that exploit the commonalities among the dialects. One research thrust involves developing the Pan-Arab Morphological Analyzer (PAMA). Initially, we will examine it with mor- phological knowledge for MSA and Egyptian, but hope to extend it to other dialects (Levantine, Iraqi, Moroccan, Yemeni, and so on). We plan to start an initial work we have done on automatic transcription and disambiguation for MSA to the dialects, using MSA as the common language. The PAMA is being used as a core component in many different applications that approximate dialect-to-MSA translation by intentionally pe- riph- es given dialect input.

Front row, from left to right: Ana Nenкова, Dr. Becky Passonneau, Lisek Shrestha, Professor Kathy McKeown, Smara Muresan, Jen- Yuan Yeh, Elena Filatova, Dr. Barry Schiffman. Back row: Aaron Hyten, Saba Galan-Guldemond, David Eisner, Michael Galvey, Dr. Oliver Hobbs, Dr. Claire Fontaine.
**Recent & Upcoming PhD Defenses**

**Gaurav KC**
Advisor: Al Ahon
**Abstract:**
We address the significant problem of identifying and processing resources against attack mechanisms commonly utilized by malware such as Internet worms and Internet cracking tools. These attacks typically exploit security vulnerabilities summarized in updates for middleware, system libraries, and even the underlying operating system, to breach and compromise the targeted software. We analyzed the anatomy of common process-subversion attacks, and derived from it a model for detecting and defending against these defense techniques that thwart such attacks. We have incorporated this technique into our ESM tool called esBench, which can be used to perform security evaluations of defense techniques in implemented systems.

**Simon Lok**
Advisor: Steve Feiner
**Abstract:**
Layout refers to the process of determining the sizes and positions of the objects that are part of an information presentation. Automated layout refers to the use of a computer program to automate either all or part of the layout process. This field of research lies at the crossroads of artificial intelligence and computer graphics. Automated layout of presentations is becoming increasingly important as the amount of data that we need to present rapidly overtakes our ability to present it manually. We present a set of novel techniques that can assist an automated layout system to produce an effective presentation of information, given a set of components to display and metadata describing the content of these components. Unlike the vast majority of previous approaches to layout and numerous related problems, our techniques attempt to maximize the information gain in the process by a human graphic designer when manually creating a presentation.

These techniques include a secure software-execution environment to counter the problem of process-subversion attacks. These techniques also offer a potential for improving the efficiency of search by using the process-subversion attacks to automatically analyze and improve the performance of search algorithms.

**David Evans**
Advisor: Kathy McKean
**Title:** Natural Language Generation in Text Summarization

**Abstract:**
Early work in the computational treatment of natural language focused on summarization and machine translation. In my research, I have concentrated on the area of summarization of documents in different languages. This thesis presents my work on multilingual text similarity. This work enables the identification of short units of text (usually sentences) that contain similar information even though they are written in different languages and add support for other languages. An in-depth examination and evaluation of the system is performed using Arabic and English data. I also apply the concept of multilingual text similarity to summarization in two different systems. The first improves readability of English summaries of Arabic text by replacing machine translated Arabic sentences with highly similar English sentences when possible. The second is a novel summarization system that supports comparative analysis of Arabic and English documents in two ways. First, given Arabic and English documents that describe the same event, SimFinderML clusters sentences to present information that is supported by both the Arabic and English documents. Second, the system provides an analysis of how the Arabic and English documents differ by presenting information that is supported exclusively by documents written in only one language. This novel form of summarization is a first step at analyzing differences in perspectives from news reported in different languages.

**Shlomo Hershkop**
Advisor: Saul Soito
**Behavior-based Email Analysis with Application to Spam Detection**

**Abstract:**
Email is the “killer network” among the Web sites that are ubiquitous and pervasive. In a relatively short timeframe, the Web has rendered email service utterly and deeply entrenched in our modern society primarily due to the power of its communication substrate linking people and organizations around the globe. Much work on email technology has focused on making email easy to use; however, the need for a wide variety of information and information types to be conveniently, reliably, and efficiently sent throughout the Internet. However, the analysis of the back-end of an email service is much more complicated and provides an opportunity for research. The Problem: Email clients provide only partial information—users have to manage multiple mailboxes, search or prioritize large amounts of email. Our thesis is that advanced email clients can provide new opportunities for applications to increase email productivity and extract new information from email archives.

This thesis presents an imple-mented framework for data mining behavior models from email archives. The Email Mining Toolkit (EMT) is a data mining toolkit developed for the purposes of these endeavors. The EMT was designed to analyze offline email corpora, including the entire set of email sent and received by an individual user revealing much information about individual users and their behavior. Prior to groups of users in an organization. A number of machine learning and anomaly detection algorithms are embedded in the system to model the user’s email behavior in order to classify email for a variety of tasks. These techniques are successfully applied to the task of classification, spam detection, and analysis to retrieve information about user’s behavior. We organize the core functionality of the EMT into a package of tools called the Profiling Email Toolkit (PET). A novel contribution in PET is the focus on analyzing real-time email information from both an individual and an organization in a standard framework. PET includes new algorithms that combine multiple models using a variety of features extracted from email to achieve higher accuracy and precision. These two packages are the first system to be run in a professional environment.

**Researcher:**

- **The Henry and Gertrude Rothschild Chair in Computer Science** was generously endowed by Henry Rothschild and his wife Gertrude Rothschild. Gertrude is the House Professor of Sciences at Massachusetts Institute of Technology and Chair of the Fu Foundation School of Engineering. Pictured are Professor Rothschild and Professor Kathy McKean, the recipient of the chair.

- **The Natural Language Processing (NLP) research group** led by Professor Kathy McKean and supported by the Office of the President and the Office of the Provost, Columbia University in the City of New York Software Industry Association (NYSIA), testified in writing to the U.S. Senate Committee on Banking, Housing and Urban Affairs during a Hearing on “Examining the Financial Services Industry’s Responsibilities and Role in Preventing Identity Theft and Protecting Sensitive Financial Information”, mentioning Prof. Malkin’s project analyzing the security configuration of protected servers. Part of the testimony read: “Dr. Malkin and her team made a systematic study of the cryptographic strength of thousands of “secure” servers on the Internet... Dr. Malkin’s study probed 25,000 secure Web servers to determine if SSL was being properly configured and whether it was employed in the most secure way. Improper configuration can lead to attacks on servers, stolen data, identity theft, break-ins, etc. Dr. Malkin’s project is the most extensive study of actually existing server security on the Internet. The team’s findings, relevant to these hearings, included some serious weak-nesses in how Web servers, including e-commerce servers employed by financial service companies, are currently being configured. These shortcomings are quite serious, and pose risks both to the consumers and the providers in the financial services industry. Financial server security can be increased both by popularizing the correct configurations and, possibly, by greater government oversight in this area.”

- **Amelie Marian (Ph.D. ’05) started as a tenure-track assistant professor at Rutgers University, in the Computer Science Department, in Fall 2005.**

- **WINTER 2006**

- **CSIQS WINTER 2006**
On October 17, a reception was held in honor of the recent establishment of several endowed chairs in the Computer Science Department.

On October 17, a reception was held in honor of the recent establishment of several endowed chairs in the Computer Science Department. The program consisted of three technical keynote lectures from Alan Halevi (IBM, New York), Craig Neville-Manning (Google Research, Mountain View), and Michael Stonebraker (MIT), and a poster session for graduate students to present their latest research. The event was sponsored by IBM Research, with additional funding from Columbia Graduate Student Advisory Council.

Scott Brinker
Kamra
Levi Professor of Mathematical Science; and Professor Julia Hirschberg, Julian Clarence Chang Professor in Intelligence, Cognitive Science, Linguistics, Natural Language Processing, Artificial Intelligence, Intelligence, Cognitive Science, Signal Processing, Pattern Recognition, etc.)

The Machine Learning Laboratory, led by Professor Tsymbal, received two one-year awards from the KDD Program (a joint effort between the Intelligent Control Community and the NSF). The first was a KDD Program Award for a proposal on “Correspondence in Learning: Combining Domain KDD Challenges for KDD.” The second was a KDD Challenge Award, made in August, for a proposal on “Text and Author Identity as a Permutation Learning Problem.”

The National Institutes of Health (NIH) has announced that Columbia will be focusing on multi-scale study of cellular networks, MAGNet, at Columbia University on April 20-24, 2005. He coordinated news story on applications of computer vision to simulating physical systems and during the spring, gave a talk on “A new method for simulating physical systems and during the spring, gave a talk on “A new method for simulating physical systems in a time-efficient manner.”

Devi Gaili, Julian Clarence Chang Professor of Mathematical Methods and Computer Science and Dean of the School of Engineering and Applied Science, was elected as a Fellow of the American Academy of Arts and Sciences. The paper “Modeling and Managing Content Changes in Text Databases,” by Pames Panagiotis (Fall 2004 Columbia PhD graduate, now an assistant professor at NTU), Alexandre Noulas (a PhD student at UCLA), Jungho Cho (an assistant professor at UCLA), and Professor Luis Gravano, received the Best Paper Award at the 21st IEEE International Conference on Data Engineering (ICDE) conference held April 2005 in Tokyo. ICDE is a highly selective and prestigious database conference.

Professor Eitan Grinspun has won a large National Science Foundation grant to investigate how to create layouts that are highly effective, superior in both form and function to those generated by designers into this field to date. Specifically, we first focus on a common web-applica- tion scenario where the data is structured and only available through a common database service with heterogeneous access interfaces and constraints. By referring to the peculiarities of the sources and potentialities of a different kind of query execution plan. We also discuss an important XML data integration scenario where XML data comes from heterogeneous sources, and there- fore, the techniques for assembling these systems to automatically generate layouts that are highly effective, superior in both form and function to those generated by designers into this field to date.

Amelie Marian Advisor: Luis Gravano Evaluation of Top-k Queries over Semi-structured Data Abstract: Traditionally, query processing is performed by preparing a ranked, or “exact matches” for the queries. Data (e.g., relational) and semi-structured data (e.g., XML) data identify the query matches are too rigid, so we approach the queries in terms of both content and structure. Processing top-k queries effi- ciently in such a scenario is a chal- lenging, as the number of candi- date answers increases dramati- cally with the number of objects in the data. By pruning irrelevant data fragments as early as possible, our algo- rithm minimizes the number of candidate answers considered during query evaluation. In summary, this thesis discusses the general problem of ranking query answers over structured and semi-structured data, and returning the “best” objects, according to user preferences, in a time-efficient manner.

Barry Schillman Advisor: Kathy McKeown Learning to Identify New Information Abstract: In this dissertation, I analyzed the differences between end system services and network services. This study included a proposal to define a new scripting language called the Language for End System Services (LESS) specifically for end system service creation, compared with the existing Python scripting technology. LESS is simple, safe, portable, extensible, and can directly interact with users and directly control media streams. LESS has a tree-like structure, making it possible to use feature interactions among LESS scripts. In this paper, I proposed a LESS-based feature handling method by using action conflict resolutions and a tree merge algorithm. Once a potential conflict is detected, my solution can be applied to resolve the detected conflict. The tree-like structure also sug- gests the possibility of automatically generating services based on decision tree induction. Since users can easily determine which services are available and how to customize or create their own services using tree-like generating services based on

Sczhou Wo Advisor: Xiaotao Wu Ubiquitous delivery of Telephony End System Services Abstract: In Internet telephony, end systems usually have CPU and memory, so they are pro- cessing of web-applications in web browsers. This is very different from traditional communication systems, which usually suffer from hard to solve the routing and addressing problems. The enhanced LESS model is designed to be used for services more distributed, and introduce many new services. However, most of the existing end system research focuses only on services in the network, which are not well-suited for end sys- tems. Therefore, it is important to conduct research specifically on end system services, including the definition of end system serv- ices, how to develop an efficient and friendly end system service creation environment, how to handle feature interactions, and how to integrate telephony serv- ices with other Internet services, such as web, email, location-based services, and networked appliance control. In this dissertation, I analyzed the differences between end system services and network services. This study included a proposal to define a new scripting language called the Language for End System Services (LESS) specifically for end system service creation, compared with the existing Python scripting technology. LESS is simple, safe, portable, extensible, and can directly interact with users and directly control media streams. LESS has a tree-like structure, making it possible to use feature interactions among LESS scripts. In this paper, I proposed a LESS-based feature handling method by using action conflict resolutions and a tree merge algorithm. Once a potential conflict is detected, my solution can be applied to resolve the detected conflict. The tree-like structure also sug- gests the possibility of automatically generating services based on decision tree induction. Since users can easily determine which services are available and how to customize or create their own services using tree-like generating services based on
I have integrated my research on end system services into our SIP user agent—SIPc. In addition to basic SIP functions and an end system service execution environment, SIPc also supports many other Internet functions, such as service discovery, event notification, networked audio control, instant messaging, and session announcement handling. Multiple functions in SIPc can interact with each other to provide new services. In this dissertation, I will discuss different ways for multi-function integration and interactions in SIPc.

Eli Androulaki
graduated from the Polytechnic University of Catalunya (Barcelona, Spain) and a Ph.D. from Tulia University (Massachusetts, USA) both in Computer Science. She moved to the Center for Computational Learning Systems at Columbia in September 2004. Her interests are in Machine Learning and Computer Biology. She is involved in several projects in applied machine learning ranging from predicting genetic regulation in the fruitfly to predicting failures in New York City’s electrical grid.

Marta Arias
has an undergraduate degree from the Polytechnic University of Catalunya (Barcelona, Spain) and a Ph.D. from Tulia University (Massachusetts, USA) both in Computer Science. She moved to the Center for Computational Learning Systems at Columbia in September 2004. Her interests are in Machine Learning and Computer Biology. She is involved in several projects in applied machine learning ranging from predicting genetic regulation in the fruitfly to predicting failures in New York City’s electrical grid.

Omer Boyaci

He is currently working with Professor Schulzrinne’s group studying Multimedia Networking and VoIP.

Seung Geo Choi
graduated from Seoul National University with a B.S. in Computer Engineering in February 1999. He worked at Bluebird Soft Co. in Korea from 1999-2002, did an M.S. in Computer Science and Engineering at Seoul National University from 2003-2005, and started as a Ph.D. student in Fall 2005. Seun Geo’s main research interests are cryptographic protocols and applying cryptographic techniques (e.g., bilinear mapping based cryptography) to problems arising in privacy issues. He is also interested in other aspects of theoretical cryptography and the interaction of cryptography and secure systems. He is currently working with Professor Moti Yung.

Rebecca L. Collins
graduated from the University of Tennessee at Knoxville with an M.S. in Computer Science in December 2004. She began Columbia’s Ph.D. program in Fall 2005. Her research interests include design and scheduling of distributed systems, and erasure codes. She is currently working with Professor Nieth’s group studying client communications in thin client environments and with Professor Cai exploring the cycle balancing problem in latency insensitive systems. She is currently supported by a three-year NDSEG Fellowship.

Moena Diab
received her PhD in Computational Linguistics in 2003 from the University of Maryland College Park. Her work focused on solving problems of ambiguity in natural language semantics exploiting multilingual evidence. She did her postdoctoral work at Stanford University working on issues of Arabic Natural Language processing. She then joined Columbia University in Feb 2004. Her research interests include computational lexical semantics, machine learning, computational psycholinguistics and comparative linguistics. As an Associate Research Scientist at the Center for Computational Learning Systems (CCLS), Dr. Diab’s current work addresses problems of processing and modeling Arabic dialects specifically and resource poor languages generally. For more info check http://www.cs.columbia.edu/~moena.

David K. Etison
graduated from Columbia University with a B.S. in Computer Science in May 2002. He remained on campus and worked in Prof. McKeown’s Natural Language Processing Group as a programmer/analyst from 2002-2005. He was also a part-time student during this time, receiving his M.S. in Computer Science in May 2005. David’s work is in computational narratology, which synthesizes his interests in AI and creative writing. He is working on his project with prof. Schulzrinne as well as professors and students in the School of the Arts.

Jinwei Gu
graduated from Tsinghua University in China with a B.S. in June 2002 and M.S. in June 2005, both in Department of Computer Science. He was a visiting student in Microsoft Research Asia from 2003 to 2004, and started as a Ph.D. student in the Computer Science Department in Fall 2005. His main research interests are Model-based Graphics and Vision, Machine Learning, Image Analysis, and Biometrics. He is currently working with Professor Shree Nayar, Professor Peter Belhumeur and Prof. Ravin Ramamoorthi studying Bidirectional Texture Function and its application to realistic rendering.

Nizar Habash
received his PhD in 2003 from the Computer Science Department, University of Maryland College Park. He is an Associate Research Scientist at the Center for Computational Learning Systems in Columbia University. His research interests include machine translation, natural language generation, lexical semantics, and combining rule-based and corpus-based resources for natural language processing. He is currently working on computational modeling of Arabic dialects and Arabic-English Machine Translation. For more information, visit his website at http://www.NizarHabash.com.

Neeraj Kumar
graduated from Georgia Tech in May 2008 with a double degree in Computer Science and Aeronautical Engineering, and he started as a Ph.D. student in computer science at Columbia in Fall 2005. He is in the vision and graphics group working with Professor Shree Nayar on topics related to computer vision and image processing. He has interned in past summers at Microsoft Research and NVIDIA corporation.

Kevin Maetvef
visited the computer science department for the summer of 2005. While at Columbia he worked with Ronit Rubinfeld and Rocco Servedio on algorithms and lower bounds for property testing of Boolean functions. Kevin is a third year Ph.D. student in theoretical computer science at MIT.

Ronit Rubinfeld
visited the computer science department for the summer of 2005 (her second summer in a row at Columbia). While at Columbia she worked with Kevin Maetvef and Rocco Servedio on algorithms and lower bounds for property testing of Boolean functions. Ronit is a Professor of Electrical Engineering and Computer Science at MIT.

Pannagadatta K. Shivashwamy
graduated from University of Mysore with a B.E. in Computer Science in 2001. He worked at Cisco Systems and Yahoo! India before joining Indian Institute of Science in 2003 for an M.E. in Systems Science and Automation. Pannagadatta’s main research interests are in Machine Learning and its applications. He is currently working with Prof. Tony Jebara and Dr. Martin Jansche. His current work focuses on applications to power and permutation invariance in Support Vector Machines.